Subject: DRBD?
Posted by cdevidal on Wed, 12 Jul 2006 21:02:28 GMT
View Forum Message <> Reply to Message

Anyone know if I can run DRBD on a CentOS 4 host?

Subject: Re: DRBD?
Posted by aistis on Thu, 13 Jul 2006 12:05:34 GMT
View Forum Message <> Reply to Message

Should work, DRBD modules are compiled in. Just compile userland tools.

Subject: Re: DRBD?
Posted by wfischer on Thu, 13 Jul 2006 14:12:41 GMT
View Forum Message <> Reply to Message

just fyi: if you need some more information - we had a talk about clustering at German's Linuxtag, there was also a part about clustering OpenVZ with DRBD and Heartbeat.

some infos:
paper: http://www.linuxtag.org/2006/fileadmin/linuxtag/dvd/12080-pa per.pdf
slides: http://www.linuxtag.org/2006/fileadmin/linuxtag/dvd/12080-cl uster-vm.pdf
a nice pic of the presentation posted by kir in the blog
http://community.livejournal.com/openvz/6310.html

Subject: Re: DRBD?
Posted by cdevidal on Thu, 13 Jul 2006 21:05:28 GMT
View Forum Message <> Reply to Message

wfischer wrote on Thu, 13 July 2006 10:12just fyi: if you need some more information - we had a talk about clustering at German's Linuxtag, there was also a part about clustering OpenVZ with DRBD and Heartbeat.

Yeah I saw that paper, forgot about it. Thanks for reminding me. Mayhaps the author has tips and scripts and config files he could share...

Subject: Re: DRBD?
Posted by cdevidal on Fri, 14 Jul 2006 12:23:13 GMT
View Forum Message <> Reply to Message

cdevidal wrote on Thu, 13 July 2006 17:05Mayhaps the author has tips and scripts and config files he could share...

Oh, that author is you!

Do you have any tips/scripts/config files you could share?

Subject: Re: DRBD?
Posted by cdevidal on Fri, 14 Jul 2006 12:44:08 GMT
View Forum Message <> Reply to Message

aistis wrote on Thu, 13 July 2006 08:05Should work, DRBD modules are compiled in. Just compile userland tools.

I just realized what you were telling me. Holy smokes, you're right!

That means I just install the userland tools, set up a partition and run my virtual machines off of it.

Subject: Re: DRBD?
Posted by wfischer on Fri, 14 Jul 2006 13:10:38 GMT
View Forum Message <> Reply to Message

Hi again,

yea - just compile and install DRBD's userland tools (you can get the source on www.drbd.org). Ensure that you use the same DRBD version of the userland tools as the DRBD version that is included in the OpenVZ kernel (I think in the OpenVZ kernel is currently drbd 0.7.17 - but check it to be sure).

Then create a drbd device, and put the /vz filesystem on it (ensure that you run the mkfs.ext3 on the /dev/drbd0 device, not on the lower device like /dev/hda5 or so...).

In our example setup in the paper, we moved the /etc/sysconfig/vz-scripts/ directory and the /etc/sysconfig/vz file also to /vz/cluster/etc/sysconfig/vz-scripts and /vz/cluster/etc/sysconfig/vz (and put symlinks on the original locations). In that way you have the vz config also on the mirrored device.

We will setup a private OpenVZ cluster mirrored by drbd within the next two weeks - I'll try to give a complete documentation on how to setup this on the wiki if I have enough time.

best wishes,
Werner

Subject: Re: DRBD?
Posted by cdevidal on Fri, 14 Jul 2006 13:19:05 GMT
View Forum Message <> Reply to Message

wfischer wrote on Fri, 14 July 2006 09:10yea - just compile and install DRBD's userland tools (you can get the source on www.drbd.org). Ensure that you use the same DRBD version of the userland tools as the DRBD version that is included in the OpenVZ kernel (I think in the OpenVZ kernel is currently drbd 0.7.17 - but check it to be sure).

I've always used precompiled RPMs but perhaps they're not the same version number. Thanks for the warning.

I could also compile SRPMs to match...

wfischer wrote on Fri, 14 July 2006 09:10Then create a drbd device, and put the /vz filesystem on it (ensure that you run the mkfs.ext3 on the /dev/drbd0 device, not on the lower device like /dev/hda5 or so...).

In our example setup in the paper, we moved the /etc/sysconfig/vz-scripts/ directory and the /etc/sysconfig/vz file also to /vz/cluster/etc/sysconfig/vz-scripts and /vz/cluster/etc/sysconfig/vz (and put symlinks on the original locations). In that way you have the vz config also on the mirrored device.

We will setup a private OpenVZ cluster mirrored by drbd within the next two weeks - I'll try to give a complete documentation on how to setup this on the wiki if I have enough time.

Nifty!

If you don't, I will. I'm working on the same from my end on CentOS 4. I've done it with Heartbeat+DRBD+VMware, now it's time for a "real" virtualization suite

Subject: Re: DRBD?
Posted by cdevidal on Fri, 14 Jul 2006 13:27:21 GMT
View Forum Message <> Reply to Message

wfischer wrote on Fri, 14 July 2006 09:10Ensure that you use the same DRBD version of the userland tools as the DRBD version that is included in the OpenVZ kernel (I think in the OpenVZ kernel is currently drbd 0.7.17 - but check it to be sure).

That is correct.

Found this:
CentOS 4 DRBD userland tool RPMs

And this:
Other distros

Unfortunately everything is 0.7.20.

Are you certain the versions must precisely match?

Subject: Re: DRBD?
Posted by wfischer on Fri, 14 Jul 2006 13:44:43 GMT
View Forum Message <> Reply to Message

cdevidal wrote on Fri, 14 July 2006 15:27Are you certain the versions must precisely match?

not 100% sure - but I think the API-version of the userland tools and the module must be the same. See http://svn.drbd.org/drbd/branches/drbd-0.7/ChangeLog for the api version of the different drbd versions.

unfortunately 0.7.17 has api:77 and 0.7.20 has api:79.

But again, I'm not 100% sure if that is the only thing that must match. To be sure I'd recommend to really use the same version.
If you cannot find prebuilt ones, it is very easy to create rpm's out of drbd's source (you can do a "make rpm", a spec file is already included in the source)

greetings,
Werner

Subject: Re: DRBD?
Posted by cdevidal on Fri, 14 Jul 2006 14:30:00 GMT
View Forum Message <> Reply to Message

wfischer wrote on Fri, 14 July 2006 09:44If you cannot find prebuilt ones, it is very easy to create rpm's out of drbd's source (you can do a "make rpm", a spec file is already included in the source)

A spec of the userland tools?

Subject: Re: DRBD?
Posted by wfischer on Fri, 14 Jul 2006 14:39:52 GMT
View Forum Message <> Reply to Message

cdevidal wrote on Fri, 14 July 2006 16:30A spec of the userland tools?
A spec for both the userland tools and the module. Normally, when you want to compile drbd you do not have a kernel where the module is already included. So the spec is for both the userland tools and the module.

btw: maybe I'll install our cluster this weekend, so perhaps the documentation will be in the wiki soon
btw #2: we are using centos 4, too

Subject: Re: DRBD?
Posted by cdevidal on Fri, 14 Jul 2006 14:46:11 GMT
View Forum Message <> Reply to Message

wfischer wrote on Fri, 14 July 2006 10:39A spec for both the userland tools and the module. Normally, when you want to compile drbd you do not have a kernel where the module is already included. So the spec is for both the userland tools and the module.

Oh, right right right.

wfischer wrote on Fri, 14 July 2006 10:39btw: maybe I'll install our cluster this weekend, so perhaps the documentation will be in the wiki soon
btw #2: we are using centos 4, too

Nifty!

btw: drbd-0.7.17 RPMS Even more nifty!

Someone (perhaps me?) should mirror them in case the DRBD programmers decide to take them down like they did 0.7.16 and below.

Well if you don't get the docs in the Wiki I'll try to. But it won't be this weekend :-\

Subject: Re: DRBD?
Posted by wfischer on Fri, 14 Jul 2006 15:06:36 GMT
View Forum Message <> Reply to Message

cdevidal wrote on Fri, 14 July 2006 16:46btw: drbd-0.7.17 RPMS Even more nifty!

Someone (perhaps me?) should mirror them in case the DRBD programmers decide to take them down like they did 0.7.16 and below.
this repository is a support repository by linbit (the guys how develop drbd). you can browse through the repository, but cannot download any files, unless you bought a support packet (see

http://www.linbit.com/en/drbd/drbd/support/).
this is one of the ways how they make money to finance the ongoing development of drbd.

Subject: Re: DRBD?
Posted by cdevidal on Fri, 14 Jul 2006 15:10:39 GMT
View Forum Message <> Reply to Message

wfischer wrote on Fri, 14 July 2006 11:06this repository is a support repository by linbit (the guys how develop drbd). you can browse through the repository, but cannot download any files, unless you bought a support packet (see http://www.linbit.com/en/drbd/drbd/support/).
this is one of the ways how they make money to finance the ongoing development of drbd.

Ohhhhh that's right, I forgot.

Well I could use the included .spec, like you said I always prefer RPMs to tarballs and since building an SRPM is about as much work as installing a tarball (easier, actually), I'll go that route.

Subject: Re: DRBD?
Posted by wfischer on Wed, 26 Jul 2006 11:45:58 GMT
View Forum Message <> Reply to Message

Hi again,

I found postings from Lars (one of the drbd developers):
http://lists.linbit.com/pipermail/drbd-user/2006-May/005027. html and
http://lists.linbit.com/pipermail/drbd-user/2006-June/005051 .html

there he gives a short explanation on version numbers for api and proto version. As I already thought, the api version of the userland tools must match the api version of the kernel module.

I checked the drbd version of the last three openvz kernels, see below:
[root@localhost ~]# cat /proc/version
Linux version 2.6.8-022stab077.1 (root@kern268.build.sw.ru) (gcc version 3.3.3 20040412 (Red Hat Linux 3.3.3-7)) #1 Fri Apr 21 16:50:02 MSD 2006
[root@localhost ~]# modprobe drbd
[root@localhost ~]# cat /proc/drbd
version: 0.7.17 (api:77/proto:74)
SVN Revision: 2093 build by phil@mescal, 2006-03-06 15:04:12
 0: cs:Unconfigured
 1: cs:Unconfigured
[root@localhost ~]#

[root@localhost ~]# cat /proc/version
Linux version 2.6.8-022stab078.10 (root@rhel4-32) (gcc version 3.4.5 20051201 (Red Hat

3.4.5-2)) #1 Wed Jun 21 12:01:20 MSD 2006
[root@localhost ~]# modprobe drbd
[root@localhost ~]# cat /proc/drbd
version: 0.7.17 (api:77/proto:74)
SVN Revision: 2093 build by phil@mescal, 2006-03-06 15:04:12
 0: cs:Unconfigured
 1: cs:Unconfigured
[root@localhost ~]#

[root@localhost ~]# cat /proc/version
Linux version 2.6.8-022stab078.14 (root@kern268.build.sw.ru) (gcc version 3.3.3 20040412 (Red Hat Linux 3.3.3-7)) #1 Wed Jul 19 16:02:34 MSD 2006
[root@localhost ~]# modprobe drbd
[root@localhost ~]# cat /proc/drbd
version: 0.7.20 (api:79/proto:74)
SVN Revision: 2260 build by phil@mescal, 2006-07-04 15:18:57
 0: cs:Unconfigured
 1: cs:Unconfigured
[root@localhost ~]#

Regarding instructions building a drbd/heartbeat cluster with openvz for the openvz wiki: finally today I have time to do this - so in about 12 hours the infos about that should be in the wiki. [update: unfortunately I won't finish today - I have my two CentOS boxes now running for this, but it takes a little longer than I thought - I post here once the content is in the wiki]

best wishes from Austria,
Werner

Subject: Re: DRBD?
Posted by wfischer on Sun, 30 Jul 2006 18:21:33 GMT
View Forum Message <> Reply to Message

I have documented the first part of how to use DRBD with OpenVZ in http://wiki.openvz.org/HA_cluster_with_DRBD_and_Heartbeat

The info on setting up Heartbeat and how to do updates (especially how to do OpenVZ kernel updates that contain a new version of DRBD, which is a little tricky) will follow soon.

best regards,
Werner

Subject: Re: DRBD?
Posted by cdevidal on Mon, 31 Jul 2006 17:39:15 GMT

wfischer wrote on Sun, 30 July 2006 14:21I have documented the first part of how to use DRBD with OpenVZ

Woohoo! I've had to pause my work, so I'm glad you've got something up there.

wfischer wrote on Sun, 30 July 2006 14:21The info on (...) how to do updates (especially how to do OpenVZ kernel updates that contain a new version of DRBD, which is a little tricky) will follow soon.

Wonderful! That info, plus knowing which OpenVZ files to copy to the DRBD partition were my unknowns.

Question: Why not two DRBD partitions, one on each node, and run a handful of VPSes on the first node and a handful on the second, so that the second node's CPU cycles and RAM are not sitting idle? Or were you just trying to keep things simple?

If you do such a setup, DRBD's Group parameter is very helpful when you have two DRBD devices on one hard drive. The first group synchronizes, then the second, but not in parallel (as would be the case if you had two drives). Set one DRBD device in one group and the other in a new group.

Subject: Re: DRBD?
Posted by wfischer on Tue, 01 Aug 2006 08:24:56 GMT

cdevidal wrote on Mon, 31 July 2006 19:39Question: Why not two DRBD partitions, one on each node, and run a handful of VPSes on the first node and a handful on the second, so that the second node's CPU cycles and RAM are not sitting idle? Or were you just trying to keep things simple?
Yea, the first reason is that the setup is more simple. And the more simple the setup is, the higher the availability will be.
The second reason is that in a active-passive configuration you can get aware of performance bottlenecks soon enough. We had a for example a cluster running, that ran Apache on node1 and MySQL on node2 (without any virtualization). When we started the project, every machine had 1,5 GB RAM. Apache needed about 500 MB, and also MySQL needed about 500 MB. After some time we discovered that Apache now needs 1 GB, and also MySQL consumes 1 GB of RAM - so if a failover would have happened the remaining cluster node would have started swapping and get very slow (in fact so slow, that it would have seemed that the cluster is down)
When you run all services on only one node, you can sooner discover those performance bottlenecks (actually before a failover happens) - and enlarge e.g. RAM like in this case.
Evan Marcus and Hal Stern have a very interesting discussion about why to use active/passive and what to answer to management when they ask: "how can I use the standby server?" You can find it in their book "Blueprints for High Availability", 2nd edition, page 417 - 425 (see

http://www.amazon.com/gp/product/0471430269/).

cdevidal wrote on Mon, 31 July 2006 19:39If you do such a setup, DRBD's Group parameter is very helpful when you have two DRBD devices on one hard drive. The first group synchronizes, then the second, but not in parallel (as would be the case if you had two drives). Set one DRBD device in one group and the other in a new group.
Yea, you are absolutely right! If someone really wants active/active, the DRBD group parameter is very valueable.

Subject: Re: DRBD?
Posted by cdevidal on Tue, 01 Aug 2006 12:40:36 GMT
View Forum Message <> Reply to Message

wfischer wrote on Tue, 01 August 2006 04:24The second reason is that in a active-passive configuration you can get aware of performance bottlenecks soon enough. We had a for example a cluster running, that ran Apache on node1 and MySQL on node2 (without any virtualization). When we started the project, every machine had 1,5 GB RAM. Apache needed about 500 MB, and also MySQL needed about 500 MB. After some time we discovered that Apache now needs 1 GB, and also MySQL consumes 1 GB of RAM - so if a failover would have happened the remaining cluster node would have started swapping and get very slow (in fact so slow, that it would have seemed that the cluster is down)
When you run all services on only one node, you can sooner discover those performance bottlenecks (actually before a failover happens) - and enlarge e.g. RAM like in this case.
Evan Marcus and Hal Stern have a very interesting discussion about why to use active/passive and what to answer to management when they ask: "how can I use the standby server?"

I'm glad I talked to someone with experience

Good info, thank you.

Two questions:
1.) Did you actually perform a failover and observe it to be so slow because it was swapping?
2.) So then am I to understand that in theory load balancing and high availability aren't in conflict but in practice they are? For example OpenSSI, which gives you high availability + load balancing, but if you just have two nodes and every service fails over to the first node it gets to be so slow you might as well not have anything at all.

In other words, are load balancing and high availability mutually exclusive not in theory but in practice, at least for two nodes?

Subject: Re: DRBD?
Posted by wfischer on Fri, 04 Aug 2006 09:55:32 GMT
View Forum Message <> Reply to Message

cdevidal wrote on Tue, 01 August 2006 14:401.) Did you actually perform a failover and observe it

to be so slow because it was swapping?
Yes, I had one situation like that. As far as I remember I also saw a situation where the OOM killer finally got active, as also the sum of physical RAM+swap was not big enough.
cdevidal wrote on Tue, 01 August 2006 14:402.) So then am I to understand that in theory load balancing and high availability aren't in conflict but in practice they are? For example OpenSSI, which gives you high availability + load balancing, but if you just have two nodes and every service fails over to the first node it gets to be so slow you might as well not have anything at all. In other words, are load balancing and high availability mutually exclusive not in theory but in practice, at least for two nodes?
I have no experience with OpenSSI - I only know that it provides a single system image across many machines. When you need load balancing (like a webserver farm), you also need two clustered load balancer boxes (otherwise the load balancer would be a single point of failure). So with load balancing you need at least four machines (two load balancers and two servers) to also get high availability.

Up untill now I have not implemented a load balancing cluster yet (I only took a deeper look on linux virtual server).

another info: I updated http://wiki.openvz.org/HA_cluster_with_DRBD_and_Heartbeat - I think it is complete now. I hope I have not overlooked errors in the document, as it is rather long meanwhile.

best wishes,
Werner

Subject: Re: DRBD?
Posted by cdevidal on Fri, 04 Aug 2006 10:25:50 GMT
View Forum Message <> Reply to Message

wfischer wrote on Fri, 04 August 2006 05:55I have no experience with OpenSSI - I only know that it provides a single system image across many machines.

Oh no I was just using it as a "for example."

I was just asking if you thought that all load balancing+high availability solutions are mutually exclusive.

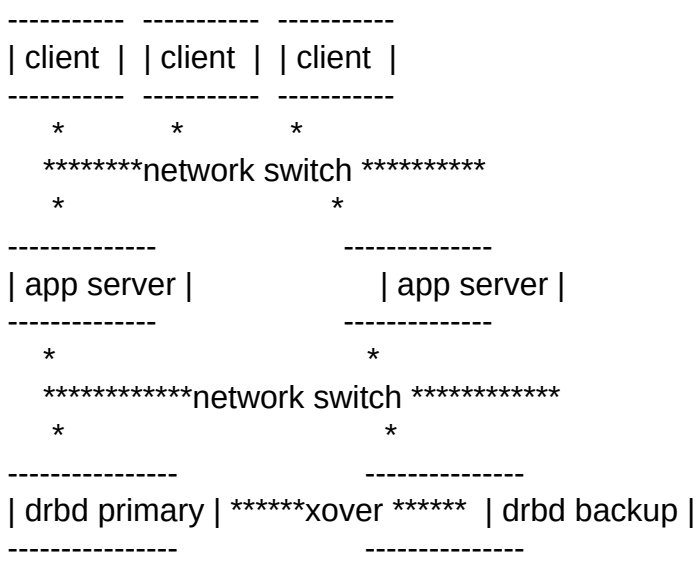
Subject: Re: DRBD?
Posted by cdevidal on Sat, 05 Aug 2006 14:03:06 GMT
View Forum Message <> Reply to Message

wfischer wrote on Fri, 04 August 2006 05:55another info: I updated http://wiki.openvz.org/HA_cluster_with_DRBD_and_Heartbeat - I think it is complete now. I hope I have not overlooked errors in the document, as it is rather long meanwhile.

Rock on! If I encounter any, I'll fix it. Wiki is great, isn't it??

Subject: Re: DRBD?
Posted by jimcooncat on Wed, 30 Aug 2006 13:08:15 GMT
View Forum Message <> Reply to Message

Trying to figure out how having both openvz and drbd on a two-node server set would be useful. In my half-baked scenario, I'm picturing a thin client setup like this:



My thought was to have OpenVZ on the "app servers" to run freenx servers, and be able to load balance/migrate sessions between the two "app servers".

Would loading openvz right on the drbd machines be able to eliminate my "app server" layer? Or am I barking up the wrong tree here?